ELZ-001

APPLICATION

FOR

UNITED STATES LETTERS PATENT

SPECIFICATION

TO ALL WHOM IT MAY CONCERN:

Be it known that John P. Kroeker, a U.S. citizen, residing in Hamilton, MA, has invented certain improvements in A NOVEL APPROACH TO SPEECH RECOGNITION of which the following description in connection with the accompanying drawings is a specification, like reference characters on the drawings indicating like parts in the several figures.

A NOVEL APPROACH TO SPEECH RECOGNITION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is related to the following U.S. patent applications, of common assignee, from which priority is claimed, and the contents of which are incorporated herein in their entirety by reference:

"A Novel Approach to Speech Recognition," U.S. Provisional Patent Application Serial No. 60/192,090;

"Combined Syntactic And Semantic Search, Parsing, And Application Access," U.S. Provisional Patent Application Serial Number 60/192,091;

"Remote Server Object Architecture For Speech Recognition," U.S. Provisional Patent Application Serial Number 60/192,076; and,

"Speech Recognition Application Technology Using Web, Scripting, And Semantic Objects," U.S. Provisional Patent Application Serial Number 60/191,915.

[0002] This application is also related to the following copending U.S. patent applications, the contents of which are incorporated herein in their entirety by reference:

"Phonetic Data Processing System and Method," U.S. Patent Application Serial Number _____, attorney docket number ELZ-2;

"Remote Server Object Architecture For Speech Recognition," U.S. Patent Application Serial Number _____, attorney docket number ELZ-3; and,

"Web-Based Speech Recognition With Scripting and Semantic Objects," U.S. Patent Application Serial Number _____, attorney docket number ELZ-4.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0003] Not Applicable

REFERENCE TO MICROFICHE APPENDIX

[0004] Not Applicable

BACKGROUND OF THE INVENTION

[0005] The present invention relates to speech recognition, and more particularly, to systems for and methods of transforming an acoustical speech signal into a linguistic stream of phonetics, words and other speech components.

[0006] In general, speech recognition is a multi-layered process. Typically, a speech recognition system analyzes a raw acoustic waveform over time, and applies some complex algorithm to extract a stream of linguistic units (e.g., the phonetics, words, etc.). The term "stream" may also be referred to herein as a "sequence" or a "series," and the term "linguistic units" may also be referred to herein as "phonetic estimates." The term "acoustic waveform" may also be referred to herein as "acoustic signal," "audio signal," or "audio waveform." A speech recognition system may further apply various sources of linguistic constraints, so that the utterance may be finally interpreted within a practical context.

[0007] Of all of the processes and associated technologies used for speech recognition, the transformation of an acoustic signal to a linguistic stream has been the most difficult, and remains the technology gatekeeper for practical applications. The problem is essentially one of pattern recognition, and shares many of the challenges of handwriting recognition, OCR and other visual recognition technologies. The process that transforms an acoustic signal to a linguistic stream is referred to herein as the "core speech recognizer".

[0008] There have been three primary strategies for approaching the problem of realizing the core speech recognizer: (1) the statistical approach, (2) the feature approach and (3) the perceptual or bio-modeling approach. Each approach is summarized below.

(1) Statistical Recognition

[0009] The statistical recognition approach involves first reducing the incoming data stream to its essential, most basic components, then applying algorithms to examine thousands, or in some cases millions, of statistical hypotheses to find the most likely spoken word-string. The

framework used area most commonly (and nearly universally) is known as Hidden Markov Modeling (hereinafter referred to as "HMM").

(2) Recognition by Linguistic Features

[0010] This approach is based on the idea that the study of linguistics has accumulated a vast body of knowledge about the acoustic features that correspond to the phonetics of human language. Once these features are characterized and estimated, a system can integrate them statistically to derive the best guess as to the underlying spoken utterance.

[0011] The feature approach has not been very successful. However, the Jupiter system at MIT has successfully combined the statistical method with a feature-based front end. While this class of recognition system remains in an experimental stage, it performs well in limited domains.

(3) Biomodeling human perception: Partial Approaches

[0012] Humans are the only example we have of a working, efficient speech recognizer. Thus, it makes sense to try to mimic how the human brain recognizes speech. This "bio-modeling" approach may be the most challenging, as there is no definitive scientific knowledge for how humans recognize speech.

[0013] One approach to bio-modeling has been to use what is known about the inner ear, and design preprocessors based on physiological analogs. The preprocessors may be used to modify the raw acoustic signal to form a modified signal. The preprocessors may then provide the modified signal into standard pattern recognizers. This approach has yielded some limited success, primarily with regard to noise immunity.

[0014] Artificial Neural Nets (hereinafter referred to as "ANNs") fit somewhat into this category as well. ANNs have become a significant field of research, and provide a class of pattern recognition algorithms that have been applied to a growing set of problems. ANNs emphasize the enormous connectivity that is found in the brain.

HMM: The Standard Prior Art Technology

[0015] The essence of the HMM idea is to assume that speech is ideally a sequence of particular and discrete states, but that the incoming raw acoustic data provides only a distorted and fuzzy representation of these pristine states. Hence the word "hidden" in "Hidden Markov Modeling." For example, we know that speech is a series of discrete words, but the representation of that speech within the acoustic signal may be corrupted by noise, or the words may not have been clearly spoken.

[0016] Speech comprises a collection of phrases, each phrase includes a series of words, and each word includes components called phonemes, which are the consonants and vowels. Thus, a hierarchy of states may be used to describe speech. At the lowest level, for the smallest linguistic unit chosen, the sub-states are the actual acoustic data. Thus, if an HMM system builds up the most likely representation of the speech from bottom to top, each sub-part or super-part helping to improve the probabilities of the others, the system should be able to just read off the word and phrase content at the top level.

[0017] The real incoming acoustic signal is continuous, however, and does not exist in discrete states. The first solution to this problem was to use a clustering algorithm to find some reasonable states that encompass the range of input signals, and assign a given datum to the nearest one. This was called VQ, or Vector Quantization. VQ worked, to a limited extent, but it turned out to be much better to assign only a probability that the given datum belonged to a state, and it might perhaps belong to some other state or states, with some probability. This algorithm goes by the name of Continuous Density HMM.

[0018] Continuous Density HMM is now the most widely used algorithm. There are many choices for how to implement this algorithm, and a particular implementation may utilize any number of preprocessors, and may be embedded into a complex system.

[0019] The HMM approach allows a large latitude for choosing states and hierarchies of states. There is a design trade-off between using phonetics or words as the base level. Words are less flexible, require more training data, and are context dependent, but they can be much more accurate. Phonetics allows either a large vocabulary or sets of small and dynamic vocabularies.

There is also a trade-off between speaker-dependent (i.e., speaker-adaptive) systems, which are appropriate for dictation, and speaker-independent systems, which are required for telephone transactions. Since individuals speak differently, HMM needs to use a large number of states to reflect the variation in the way words are spoken across the user population. A disadvantage to prior art systems that use HMM is a fundamental trade-off between functionality for (1) many words or (2) many people.

Challenges to Automatic Speech Recognition (ASR)

[0020] A publicly accessible recognition system must maintain its accuracy for a high percentage of the user population.

> "Human adaptation to different speakers, speaking styles, speaking rates, etc., is almost momentarily [i.e. instantaneous]. However, most so-called adaptive speech recognizers need sizable chunks of speech to adapt." (Pols, Louis C.W., *Flexible, robust, and efficient human speech recognition*, Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21 (1997), 1-10)

Variation in users includes age, gender, accent, dialect, behavior, motivation, and conversational strategy.

[0021] A publicly accessible speech recognition system must also be robust with respect to variations in the acoustical environment. One definition of environmental robustness of speech recognition is maintaining a high level of recognition accuracy in difficult and dynamically-varying acoustical environments. For telephone transactions, variations in the acoustical environment may be caused by variations in the telephone itself, the transmission of the voice over the physical media, and the background acoustical environment of the user.

> "Natural, hands-free interaction with computers is currently one of the great unfulfilled promises of automatic speech recognition (ASR), in part because ASR systems cannot reliably recognize speech under everyday, reverberant conditions that pose no problems for most human listeners." (Brian E. D. Kingsbury, *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*, PhD thesis, UC Berkeley, 1998.)

[0022] In many respects, adverse effects on the acoustic signal to be recognized are getting worse with new communications technology. Speaker-phone use is becoming more common, which increases the noise and the effect of room acoustics on the signal. The speech signal may be degraded by radio transmission on portable or cellular phones. Speech compression on wire-line and cellular networks, and increasingly, on IP-telephony (i.e., voice-over-IP), also degrades the signal. Other sources of background noise include noise in the car, office noise, other people talking, and TV and radio.

> "One of the key challenges in ASR research is the sensitivity of ASR systems to real-world levels of acoustic interference in the speech input. Ideally, a machine recognition system's accuracy should degrade in the presence of acoustic interference in the same way a human listener's would: gradually, gracefully and predictably. This is not true in practice. Tests on different state-of-the-art ASR systems carried out over a broad range of different vocabularies and acoustic conditions show that automatic recognizers typically commit at least ten times more errors than human listeners." (Brian E. D. Kingsbury, *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*, PhD thesis, UC Berkeley, 1998.)

> "While a lot of progress has been made during the last years in the field of Automatic Speech recognition (ASR), one of the main remaining problems is that of robustness. Typically, state-of-the-art ASR systems work very efficiently in well-defined environments, e.g. for clean speech or known noise conditions. However, their performance degrades drastically under different conditions. Many approaches have been developed to circumvent this problem, ranging from noise cancellation to system adaptation techniques." (K. Weber. *Multiple time scale feature combination towards robust speech recognition.* Konvens, 5. Konferenz zur Verarbeitung naturlicher Sprache, (to appear), 2000. IDIAP{RR 00-22 7})

Changes Needed to Optimize ASR

[0023] The ability of an ASR to integrate information on many time scales may be important.

> "Evidence from psychoacoustics and phonology suggests that humans use the syllable as a basic perceptual unit. Nonetheless, the explicit use of such long time-span units is comparatively unusual in automatic speech recognition systems for English." (S. L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg, *Incorporating information from*

syllable-length time scales into automatic speech recognition, ICASSP, pages 721--724, 1998.)

[0024] The ability to generalize to new conditions of distortion and noise would be of great importance:

> "The recognition accuracy of current automatic speech recognition (ASR) systems deteriorates in the presence of signal distortions caused by the background noise and the transmission channel. Improvement in the recognition accuracy in such environments is usually obtained by re-training the systems or adaptation with data from the new testing environment." (S. Sharma, *Multi-Stream Approach To Robust Speech Recognition*, OGI Ph.D. Thesis, April 1999, Portland, USA.)

[0025] It may be important to integrate information from many different aspects or features of the acoustic signal:

> "One of the biggest distinctions between machine recognition and human perception, is the flexible multi-feature approach taken by humans versus the fixed and limited feature approach by pattern recognition machines." (Pols, Louis C.W., *Flexible, robust, and efficient human speech recognition*, Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21 (1997), 1-10.)

[0026] Or again:

> "Human listeners generally do not rely on one or a few properties of a specific speech signal only, but use various features that can be partly absent ('trading relations'), a speech recognizer generally is not that flexible. Humans can also quickly adapt to new conditions, like a variable speaking rate, telephone quality speech, or somebody having a cold, using pipe speech, or having a heavy accent. This implies that our internal references apparently are not fixed, as they are in most recognizers, but are highly adaptive." (Pols, Louis C.W., *Flexible, robust, and efficient human speech recognition*, Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21 (1997), 1-10.)

> "However, if progress is to be made against the remaining difficult problems [of ASR], new approaches will most likely be necessary." (Herve Bourlard, Hynek Hermansky, Nelson Morgan, *Towards increasing speech recognition error rates*, Speech Communication 18, pp.205--231, 1996.)

[0027] It is an object of the present invention to substantially overcome the above-identified

disadvantages and drawbacks of the prior art.

SUMMARY OF THE INVENTION

[0028] The present invention is based on the concept that speech is an acoustical signal encoded with information, and the human brain applies a set of rules to the encoded signal to decode the information. Once those rules are determined, an artificial system can utilize the rules to similarly decode the signal and extract the information.

[0029] The essential principle of speech, we believe, is that the human brain hears in a highly parallel, multi-faceted fashion, and performs complex transformations at many levels. The stability, generality, and environmental robustness of such front-end processing leads to the qualities and features that distinguish this approach from other approaches to ASR.. The technology described and claimed herein is strongly driven by the neurophysiology of human perception. A mathematical model of neural functioning is constructed and arranged to map the way the higher brain processes speech. This approach goes beyond the preprocessing steps, and avoids the limitations of ANNs. Like the brain, this mathematical model is a highly parallel series of processes, each of which performs specific functions, and which, taken together, take speech apart and put it back together in an intricate structure that builds in error-correction, robustness and the ability to generalize to new conditions.

[0030] The advantages have proven to be robustness and generality. Just as biomodeling the speech preprocessing provides some robustness, biomodeling the entire recognizer provides significantly more robustness.

[0031] The approach used in the present invention to the core recognition problem sidesteps the fundamental assumptions of HMM based technologies. There are no assumptions of discrete states for the acoustic data. Thus, noise or less probable utterances cannot miscue one data state for another.

[0032] Furthermore, HMM uses a web of phonetic hypotheses that depend critically on context. The phonetic front end of the present invention produces definite phonetic signals. This

provides a context independence that is critical to performing well in real-world applications where the acoustic context and linguistic context are extremely variable.

[0033] Note that the Jupiter system mentioned above, while avoiding some of the difficulties of the HMM approach with a feature-estimation front-end, shares the statistical back-end of the HMM approach. With both front and back ends depending strongly on context, complexity grows, and success in one domain becomes difficult to translate to others. A key difference in the present invention is that the phonetic recognizer provides a feed-forward, context-independent stream of phonetic estimates. This allows simplification of follow-on processing steps.

[0034] The foregoing and other objects are achieved by the invention which in one aspect comprises a speech recognition system for transforming an acoustic signal into a stream of phonetic estimates. The system includes a frequency analyzer for receiving the acoustic signal and producing as an output a short-time frequency representation of the acoustic signal. The system further includes a novelty processor that receives the short-time frequency representation of the acoustic signal, and separates one or more background components of the representation from one or more region of interest components of the representation. The novelty processor produces a novelty output that includes the region of interest components of the representation according to one or more novelty parameters. The system also includes an attention processor that receives the novelty output and produces a gating signal as a predetermined function of the novelty output according to one or more attention parameters. The system further includes a coincidence processor that receives the novelty output and the gating signal, and produces a coincidence output. The coincidence output includes information regarding co-occurrences between samples of the novelty output over time and frequency. The coincidence processor selectively gates the coincidence output as a predetermined function of the gating signal, so as to produce a gated coincidence output according to one or more coincidence parameters. The system also includes a vector pattern recognizer and a probability processor for receiving the gated coincidence output and producing a phonetic estimate stream representative of acoustic

signal.

[0035] In another embodiment of the invention, the short-time frequency representation of the audio signal includes a series of consecutive time instances. Each consecutive pair is separated by a sampling interval, and each of the time instances further includes a series of discrete Fourier transform (DFT) points, such that the short-time frequency representation of the audio signal includes a series of DFT points.

[0036] In another embodiment of the invention, for each DFT point, the novelty processor calculates a first average value across a first predetermined frequency range and a first predetermined time span. The novelty processor also calculates a second average value across a second predetermined frequency range and a second predetermined time span. The novelty processor then subtracts the second average value from the first average value so as to produce the novelty output.

[0037] In another embodiment of the invention, the first frequency range, the first time span, the second frequency range and the second time span are each a function of one or more of the novelty parameters.

[0038] In another embodiment of the invention, the first predetermined frequency range is substantially centered about a frequency corresponding to DFT point, and the first predetermined time span is substantially centered about an instant in time corresponding to the DFT point.

[0039] In another embodiment of the invention, the first predetermined frequency range is substantially smaller than the second predetermined frequency range.

[0040] In another embodiment of the invention, the first predetermined time span is substantially smaller than the second predetermined time span.

[0041] In another embodiment of the invention, the second predetermined time span is large relative to the second predetermined frequency range.

[0042] In another embodiment of the invention, the second predetermined frequency range is large relative to the second predetermined time span.

[0043] In another embodiment of the invention, for each DFT point, the novelty processor

further calculates one or more additional novelty outputs. Each additional novelty output is defined by characteristics including a distinct first frequency range, first time span, second frequency range and second time span, each characteristic being a function of one or more of the novelty parameters.

[0044] In another embodiment of the invention, the coincidence output includes a sum of products of novelty output points over two sets of novelty output points.

[0045] In another embodiment of the invention, the two sets of novelty output points includes a first set of novelty output points corresponding to a first instant in time and a second set of novelty output points corresponding to a second time instance.

[0046] In another embodiment of the invention, the two sets of novelty output points all correspond to a single time instance.

[0047] In another embodiment of the invention, the coincidence processor performs the sum of products of novelty output points over two sets of novelty output points according to one or more selectably variable coincidence parameters including (but not limited to) time duration, frequency extent, base time, base frequency, delta time, delta frequency, and combinations thereof.

[0048] In another embodiment of the invention, each of the time instances further includes an energy value in addition to the series of novelty output points.

[0049] In another embodiment of the invention, the attention processor compares the energy value to a predetermined threshold value according to a comparison criterion, so as to produce an energy threshold determination. The attention processor then produces the gating signal as a predetermined function of the threshold determination.

[0050] In another embodiment of the invention, the one or more attention parameters include the predetermined threshold value, the comparison criterion and the predetermined function of the threshold determination.

[0051] In another embodiment of the invention, the novelty parameters, the attention parameters and the coincidence parameters are selected via a genetic algorithm.

[0052] In another aspect, the invention comprises a speech recognition system for transforming a short-time frequency representation of an acoustic signal into a stream of coincidence vectors. The system includes a novelty processor that receives the short-time frequency representation of the audio signal, and separates one or more background components of the signal from one or more region of interest components of the signal. The novelty processor also produces a novelty output including the region of interest components of the signal according to one or more novelty parameters. The system also includes a coincidence processor that receives the novelty output and the gating signal, and produces a coincidence vectors that includes data describing correlations between samples of the novelty output over time and frequency.

[0053] Another embodiment of the invention further includes an attention processor for receiving the novelty output and producing a gating signal as a predetermined function of the novelty output according to one or more attention parameters. The coincidence output is selectively gated as a predetermined function of the gating signal, so as to produce a gated coincidence output according to one or more coincidence parameters.

[0054] In another aspect, the invention comprises a method of transforming an acoustic signal into a stream of phonetic estimates. The method includes receiving the acoustic signal and producing a short-time frequency representation of the acoustic signal. The method further includes separating one or more background components of the representation from one or more region of interest components of the representation, and producing a novelty output including the region of interest components of the representation according to one or more novelty parameters. The method also includes producing a gating signal as a predetermined function of the novelty output according to one or more attention parameters. The method further includes producing a coincidence output that includes correlations between samples of the novelty output over time and frequency. The coincidence output is selectively gated as a predetermined function of the gating signal, so as to produce a gated coincidence output according to one or more coincidence parameters. The method also includes producing a phonetic estimate stream representative of

acoustic signal as a function of the gated coincidence output.

[0055] In another embodiment of the invention, the method further includes calculating a first average value across a first predetermined frequency range and a first predetermined time span. The method further includes calculating a second average value across a second predetermined frequency range and a second predetermined time span, and subtracting the second average value from the first average value so as to produce the novelty output.

[0056] In another embodiment of the invention, the method further includes calculating, for each of a plurality of DFT points from the a short-time frequency representation of the acoustic signal, one or more additional novelty outputs. Each additional novelty output is defined by characteristics including a distinct first frequency range, first time span, second frequency range and second time span, each characteristic being a function of one or more of the novelty parameters.

[0057] In another embodiment of the invention, the method further includes performing a sum of products of novelty outputs over two sets of novelty outputs according to one or more selectably variable coincidence parameters. The parameters include (but are not limited to) time duration, frequency extent, base time, base frequency, delta time, delta frequency, and combinations thereof.

[0058] In another embodiment of the invention, the method further includes comparing the energy value to a predetermined threshold value according to a comparison criterion, so as to produce an energy threshold determination, and (ii) producing the gating signal as a predetermined function of the threshold determination

[0059] In another embodiment of the invention, the method further includes selecting the novelty parameters, the attention parameters and the coincidence parameters via a genetic algorithm.

BRIEF DESCRIPTION OF DRAWINGS

[0060] The foregoing and other objects of this invention, the various features thereof, as well as the invention itself, may be more fully understood from the following description, when read together with the accompanying drawings in which:

[0061] FIG. 1 provides an overview, in block diagram form, of the complete phonetic recognition system according to the present invention;

[0062] FIG. 2 shows the center-surround receptive field novelty processing performed by the novelty processor of the system shown in FIG. 1;

[0063] FIG. 3 shows the data flow from the short-time frequency analyzer through the novelty processor for one preferred embodiment of the invention shown in FIG. 1;

[0064] FIG. 4 shows the novelty parameters produced by the GA for fricatives;

[0065] FIG. 5 shows the coincidence processing parameters and modules produced by the GA for fricatives;

[0066] FIG. 6 shows the novelty parameters produced by the GA for vowels;

[0067] FIG. 7 shows the coincidence processing parameters and modules produced by the GA for vowels;

[0068] FIG. 8 shows the novelty parameters produced by the GA for non-fricatives;

[0069] FIG. 9 shows the coincidence processing parameters and modules produced by the GA for non-fricatives;

[0070] FIG. 10 shows the specific data flow used in a preferred embodiment of the invention shown in FIG.1;

[0071] FIGs. 11A and 11B together show the high-level object-flow specification of the complete processing flow of the one preferred embodiment of the system;

[0072] FIG. 12 shows a list of explanations for the modules, functions and parameters referred to in the object flow of FIGs. 11A and 11B;

[0073] FIGs. 13A and 13B together show the ScaleMean software module used in the NoveltyRT software module from one preferred embodiment of the present invention;

[0074] FIGs. 14A, 14B, 14C, 14D, 14E, 14F and 14G together show the NoveltyRT

software module from one preferred embodiment of the present invention; and,

[0075] FIGs. 15A, 15B, 15C, 15D, 15E, 15F, 15G, 15H and 15I together show the coincidenceRT and eTrigger software modules from one preferred embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0076] FIG. 1 provides an overview, in block diagram form, of the complete phonetic recognition system 100 according to the present invention. In the system 100, a short-time frequency analyzer 102 receives a raw acoustic signal 104 and produces a short-time representation 106 of the acoustic signal. A novelty processor 108 receives the short-time representation 106 of the acoustic signal and produces a novelty output 110. An attention processor 112 receives the novelty output 110 and produces an attention gate 114. A coincidence processor 116 receives the novelty output 110 and the attention gate 114 and produces a gated coincidence output 118. A vector pattern recognizer 120 and a Bayes probabilities processor 122 further process the gated coincidence output 118, so as to produce stream of phonetic estimates 124 corresponding to the acoustic signal 104.

[0077] The Short-time Frequency Analyzer 102, the Vector Pattern Recognizer 120 and the Bayes Probabilities processor 122 are described in detail in U.S. Patents 5,027,408, 5,168,524, and 5,369,726, which are hereby incorporated by reference in their entirety. The description herein is primarily concerned with the novelty processor 108, the attention processor 112, and the coincidence processor 116, and how these components operate within the phonetic recognition system 100 as shown in FIG. 1

[0078] The phonetic recognition system 100 produces a context-independent, feed-forward phonetic stream corresponding to the acoustic signal 104. As used herein, the term "context independent" means that the data produced from any one component of the phonetic recognition system 100 can be interpreted correctly in any context, i.e., there is no need to ever go back and "re-work" the data. As used herein, the term "feed forward" means that the data flow within the

phonetic recognition system 100 is in the forward direction only.

[0079] The context-independent phonetic stream simplifies and generalizes the functions of follow-on language search and understanding functions. This capability also provides structural and system advantages, such as the ability to separate and encapsulate **modular** high-level functions. In addition, these functions are separate from the system. This, in turn, provides the capability to use different implementations, technologies, hardware and system configurations, all within the same system architecture. The high-level functions remain independent of the implementation details associated with specific sub-functions and algorithms. This independence provides a fast and proven development environment, such as:

1. Re-use of common system components allows a maximum of tested, reliable software to be used.
2. Separable functions allows modular components to be completed and tested in isolation.
3. Maintenance and support are simplified.

[0080] The short-time frequency analyzer 102 processes the acoustic signal 104 to produce a short-time frequency representation. The details are essentially the same as is described in U.S. Patents 5,027,408, 5,168,524, and 5,369,726. The time interval used is 12 msec, the input sampling rate is 8000 Hz and the length of the discrete Fourier transform (hereinafter "DFT") is 128. The 64 point DFT is averaged in bins to reduce its resolution to 40 points. Adding band-limited energy yields a stream of 41 point vectors every 12 msec. This frequency-time data stream is referred to herein in vector notation as **x**, or equivalently in functional notation as x(t, f). No other special effort is made to data-reduce the signal. Other frequency-time processing techniques (e.g., the Wigner-Ville transform) have been used in alternative embodiments, with no essential change in results. Other embodiments that perform at a higher resolution in time may be used, but at the expense of an increase in computation. The short-time frequency analyzer 102 of the present invention data-reduces less than is typical for prior-art HMM algorithms; thus the present invention preserves more information than such HMM systems.

[0081] The acoustic signal 104 is the sound produced by the speaker. The speech channel is the medium, or media, through which the sound passes before it reaches the ear of the listener. The speech signal is the result of, and the speech channel consists of, a cascade of time-varying linear or near-linear systems. The various mouthparts of the speaker can be manipulated differently, each shaping the varying exciting signal with frequency shaping effects such as resonances and anti-resonances. At any given point in time, these effects can all be approximately modeled by a cascade of linear systems, followed by a linear channel.

[0082] In the frequency domain, each of these linear systems is characterized by a Transfer Function, represented as a frequency spectrum. Because the speech systems are cascaded, one simply multiplies the source spectrum (glottal pulses or fricated noise) by all these spectra to infer the resulting spectrum. If one considers the log of the spectrum (i.e., power spectrum), then the various speech components may simply be added to infer the resulting spectrum. Next, the resultant log power spectrum may be decomposed into additive components, which will provide all there is to know about the speech signal. However, this is true only for *a single point in time.*

[0083] Each of the additive components is varying, the mouthparts are interacting in a coupled visco-elastic system, the speaker's head is moving, thus projecting the sound through a constantly changing channel, and the speaker's brain is listening and adapting and controlling all this in an unknown way. In the visual domain, the problem of perceiving a visual figure against a visual background is known as the *figure-ground problem*. A similar problem exists in the speech domain. A particular phoneme results from a particular set of muscular actions by the speaker, so for speech recognition, the object is to perceive the resulting time-varying spectrum and infer which phoneme was spoken. The figure-ground problem is that the time-varying spectrum of interest is riding on an another unknown time-varying spectrum: one of similar or possibly larger amplitude. This background consists of all the phonemes and actions taking place around the time of the phoneme of interest, as well as the changing transmission channel.

[0084] Novelty . . . newness . . . surprise; these are words that colloquially describe new information. An information processor that maximizes novelty necessarily maximizes

information. As described herein, solving the figure-ground problem in the speech domain requires separating a time varying signal of interest from a background time varying signal that has been added to it. Ideally, the confusing signal (e.g., effect resulting from the head motion of the speaker) is slower that phonetic signal of interest. In this case, one can simply subtract an average of an appropriate time-constant from the signal and so produce the phonetic signal of interest. But this is also an information-maximizing operation, in that the slow-moving signal is known, but the faster signal is relatively new. In general, the novelty processor 108 performs a differential operation, i.e., it compares two regions of the incoming short-time representation 106 of the acoustic signal. The result of the comparison is a novelty output value for each input signal sample.

[0085] The basic operation of subtracting a local background component from the signal region of interest component of the signal at a given time is referred to herein as "novelty processing," and is performed by the novelty processor 108. The novelty processor 108 performs novelty processing for each frequency (i.e., for each DFT point), at each time, so that a "novelty-processed" version of the spectral stream is the resulting novelty output. Thus, the novelty processor 108 produces a "novelty point" for each input DFT point. Since there is initially no way to know what the optimum averaging interval should be, or even whether it should change over time or frequency, the novelty processor 108 processes many different averaging intervals in parallel, via a software module referred to herein as "noveltyRT." This novelty processing is similar to processing known to occur in the brain, in both the visual and auditory systems, known as the center-surround receptive field by physiologists. The noveltyRT module processes the frequency-time data stream x(t, f) as described in the following description, and as shown in FIG. 2. At each time t, and frequency f,

1. The novelty processor 108 calculates an average of the values of the data stream x(t, f) within a center rectangle 202 to produce the value C_{AVG}. The center rectangle 202 corresponds to the region of interest component described herein. The novelty parameters that may be varied are time-lag 204 relative to time t (where t is the center of the rectangle in the time dimension), length 206 and width 208 of rectangle 202.
2. The novelty processor 108 calculates an average of the values of the data stream x(t, f)

within a surround rectangle 210 to produce the value S_{AVG}. The surround rectangle 210 corresponds to the background component described herein. The parameters used are time-lag 212 relative to time t (where t is the center of the rectangle in the time dimension),, length 214 and width 216 of rectangle 210. Note that these parameters are independent of the Center.

3. The novelty output y(t, f) at time t and frequency f is the difference of Center and a scaled version of Surround, i.e., $y(t, f) = C_{AVG} - \alpha S_{AVG}$, where α is a scaling factor greater than or equal to zero.

[0086] FIG. 3 shows the data flow from the short-time frequency analyzer 102 through the novelty processor 108 for one preferred embodiment of the invention. Note that the novelty processor 108 may produce multiple novelty outputs 110; one preferred embodiment produces six outputs 110. In FIG. 3, each set of novelty outputs 110 is shown being generated by a separate novelty instance 109 within the novelty processor. Each novelty instance 109 is characterized by a different set of novelty parameters. Each of the novelty outputs 110 is subsequently processed by several coincidence maps 220 (shown only for the first novelty instance 109). The coincidence maps 220 are processed within the coincidence processor 116, and are described in more detail herein. Further, several attention triggers 222 (also described herein) operate on each of the coincidence maps. In one preferred embodiment, there are 7 different attention triggers 222, and each coincidence map 220 may use one of these attention triggers 222.

[0087] In a preferred embodiment, the novelty processor 108 provides multiple novelty-processed versions of the original spectrum. Because each of these versions includes different and nonlinear follow-on processing, the multiple versions will result in very different and complementary views of the original signal. This redundancy contributes to robustness and generality of speech recognition. In some embodiments of the invention, the edges of the regions being compared may be weighted so that the edges tail off gradually, rather than the hard-edged regions described herein. In other embodiments, the regions being compared may have shapes other than rectangular as described herein. Some embodiments of the invention may utilize some other operation to produce the novelty output rather than the summing or averaging

described herein. Such other operations may include variance of the regions, sum of the products of points in the regions, or other mathematical operations known to those in the art.

[0088] Assuming that a given novelty version contains the time-varying signal of interest, how does one identify it? The approach described herein for the present invention is motivated by two independent views.

1. Mathematical -- use of a sum-of-products approach. This is a version of the Second order Volterra series approach discussed in the 5,027,408, 5,168,524, and 5,369,726 patents, and has advantages of relatively easy evaluation, generality, and power.

2. Empirical/physiological -- assume that there are relevant features, or sets of features that identify a particular signal. The product operation may be used as a logical conjunction (i.e., logical AND) to identify co-occurring pairs of data events, and also use summation (i.e., logical OR) to aggregate groups of co-occurring events, in order to collect large groups of data in a feasible manner.

[0089] This approach is referred to herein as "coincidence processing," because it is sensitive to the coincidence, or co-occurrence of events. The sums of products have the following 6 basic degrees of freedom, or coincidence parameters, that may be varied:

1. Time duration -- defines the size of the time span over which the coincidence processor performs the sums of products.
2. Frequency extent -- defines the frequency resolution that the coincidence processor uses for performing the sums of the products. Allows reduction in computations by reducing the overall number of frequency points used in the calculations.
3. Base time -- defines the beginning of the time sweep range.
4. Base frequency -- defines the beginning of the frequency range.
5. Delta time -- defines the amount of time between points to be multiplied. This value remains constant as the time is swept when the coincidence processor performs the sums of the products.
6. Delta frequency -- defines the frequency spacing between points to be multiplied.

The coincidence processor 116 groups like sets of these coincidence processing operations into specific groups, while varying some parameters and fixing the remaining parameters in an orderly, systematic way.

[0090] In one embodiment, coincidence processing is a sum of products over two sets of time-frequency input data. If $\mathbf{y} = y(t, f)$ is the novelty output stream 110, one output of the coincidence processor 116 is given by:

$$\text{Coincidence output} = \Sigma\, y_i\, y_j,$$

where i is the index for one predetermined set of novelty output samples , and j is the index for a second predetermined set of novelty input samples. The characteristics of each predetermined set of novelty input samples are defined by a corresponding set of coincidence parameters as described herein. The key to effective processing is in the selection and implementation of these sets, which have an enormous combinatorial potential. A distinct instance of coincidence processing (i.e., using a particular set of coincidence parameters) is referred to herein as a "coincidence map."

[0091] The particular software modules used in one preferred embodiment of the present invention are as follows:

1. eCrossColumn
2. selfAddLocalFreq
3. crossAddLocalFreq

The basic operation used by all these software modules is to sum products between two time-frequency rectangles. This operation is presented in pseudocode as follows:

```
SUMOVER2RECTANGLES( tstart, tstop, delta, f1, f2, fWidth)
      sum = 0.0;
      for ( t = tstart; t < tstop; t++ )
            for ( i = 0; i < fWidth; i++ )
                  sum += y[t+delta,f1++] * y[t, f2++ ];
      put( sum );
```

Here, the first rectangle has origin (tstart, f1), time width tstop-tstart, and frequency height fWidth. The second rectangle has an origin of (tstart+delta, f2), a width of tstop-tstart, and a height of fWidth. The operation put() places the result on the next position of the output vector.

[0092] The software module eCrossColumn is presented in pseudocode as follows:

```
eCrossColumn( delta, tstart, tstop, fWidth)
      SUMOVER2RECTANGLES( tstart, tstop, delta, 0, 0, 1)
      for ( f = 1; f <= frequencyMax-fWidth; f += fWidth )
            SUMOVER2RECTANGLES( tstart, tstop, delta, 0, f, fWidth)
```

Note that the band limited energy value is located at the zero position of the time/frequency input data (i.e., the novelty data 110) for each time t. The module eCrossColumn first calculates the limiting case of (energy * energy). Note also that the energy rectangle resulting from this calculation always has height 1. The remaining calculations include the sums of (energy * frequency sample) over consecutive frequency blocks, each fWidth wide, swept from tstart to tstop.

[0093] The software module selfAddLocalFreq is presented in pseudocode as follows:

```
selfAddLocalFreq( tstart, tstop, localN)
      for ( f1 = 1; f1 < frequencyMax-localN; f1 += localN )
        for ( f2 = 1; f2 <= f1; f2 += localN )
          SUMOVER2RECTANGLES( tstart, tstop, 0, f1, f2, localN)
```

The module selfAddLocalFreq computes the sum of products for each possible pair of blocks of frequency samples. The "self" notation indicates that the pairs occur at the same time (i.e., the "delta" argument in SUMOVER2RECTANGLES is set to zero). The size of the blocks is defined by the argument localN

[0094] The software module crossAddLocalFreq is presented in pseudocode as follows:

```
crossAddLocalFreq( delta, tstart, tstop, fWidth)
      for ( f1 = 1; f1 <= frequencyMax - fWidth; f1 += fWidth )
            for ( f2 = 1; f2 <= frequencyMax - fWidth; f2 += fWidth )
              SUMOVER2RECTANGLES( tstart, tstop, delta, f1, f2, fWidth)
```

The module crossAddLocalFreq computes the sum of products for each possible pair of blocks of frequency samples. The "cross" notation indicates that the pairs occur at the different times (i.e., the "delta" argument in SUMOVER2RECTANGLES is set to some non-zero value).

[0095] Although the embodiment of the coincidence processor described herein sums the products of novelty points over particular regions, other embodiments may use other methods of comparing and/or combining novelty points. For example, one embodiment may use the logical

"OR" of products of novelty points, while other embodiments may use the logical "AND" of pairs of novelty points. Another embodiment of the coincidence processor could simply produce a stream of novelty point combinations (e.g., products) that may be subsequently combined in different ways. Thus, the coincidence processor generally combines novelty points to detect coinciding or co-occurring events within the novelty output data stream.

[0096] The performance of the basic coincidence processing is enhanced when an appropriate "attention" gate 114 is used judiciously. The attention gate 114 forces the coincidence processor to process only those frequency samples that exist when salient events occur, such as times that coincide with an energy peak. An attention gate 114 may be expressed as a function of time a(t), which has value of "1" when a salient event occurs, and has a value of "0" otherwise. The coincidence processor may incorporate the attention gate 114 into the coincidence processing as follows:

$$\text{Coincidence output} = \sum a(t)\, y_i\, y_j ,$$

Thus, the attention gate a(t) zeros out the product (y_i y_j) except at times where a(t) = 1. Because no single attention gate is suitable for all types of coincidence processing, a preferred embodiment of the invention uses a variety of attention gates, and a particular coincidence function may used any of these, or none at all. For a given attention function, the pseudocode for the coincidence processing becomes:

```
SUMOVER2RECTANGLES-GATE( tstart, tstop, delta, f1, f2, fWidth, eGate)
       sum = 0.0;
       for ( t = tstart; t < tstop; t++ )
              if ( attention[t] )
                     for ( i = 0; i < fWidth; i++ )
                            sum += x[t+delta,f1++] * x[t, f2++ ];
       put( sum );

eCrossColumn( attention, delta, tstart, tstop, fWidth)
       SUMOVER2RECTANGLES( tstart, tstop, delta, 0, 0, 1)
       for ( f = 1; f <= frequencyMax-fWidth; f += fWidth )
              SUMOVER1RECTANGLES-GATE( tstart, tstop, delta, 0, f, fWidth)

selfAddLocalFreq( attention, tstart, tstop, localN)
       set putQuotient
       for ( f1 = 1; f1 < frequencyMax-localN; f1 += localN )
         for ( f2 = 1; f2 <= f1; f2 += localN )
           SUMOVER2RECTANGLES-GATE( tstart, tstop, 0, f1, f2, localN, attention)
```

```
crossAddLocalFreq( attention, delta, tstart, tstop, fWidth )
  for ( f1 = 1; f1 <= frequencyMax - fWidth; f1 += fWidth )
    for ( f2 = 1; f2 <= frequencyMax - fWidth; f2 += fWidth )
      SUMOVER2RECTANGLES-GATE( tstart, tstop, delta, f1, f2, fWidth, attention)
```

The main difference between these modules and the modules shown without the attention gate 114 is the gating of the sum in the sum of products function. There are also minor differences, in that the eCrossColumn module performs a simple sum of frequencies, since using the energy product with the energy gate 114 is somewhat redundant. Also, the selfAddLocalFreq module generates a sum that is normalized by the actual number of times used in the sum.

[0097] One preferred embodiment of the attention processor 112 generates seven different attention triggers as follows:

eplus

```
if ( energy[t] > 0.0 )
      attention = 1;
```

eminus

```
if ( energy[t] < 0.0 )
      attention = 1;
```

eDeltaPlus

```
if ( (energy[t]-energy[t-1] ) > 0.05 )
      attention = 1;
```

eDeltaPlusM1

```
if ( (energy[t-1] - energy[t-2] ) > 0.05 )
      attention = 1;
```

eDeltaPlusM2

```
if ( (energy[t-2] - energy[t-3] ) > 0.05 )
      attention = 1;
```

eDeltaPlusP2

```
if ( (energy[t+2] - energy[t+1] ) > 0.05 )
      attention = 1;
```

eDeltaMinus

```
if ( (energy[t]-energy[t-1] ) < -0.05 )
      attention = 1;
```

The attention parameters discussed herein are used to select which one, if any, of these attention triggers should be used to provide an attention gate 114 to the coincidence processor 116. Since any one of these triggers may be used, or none at all, there are eight attention gates possible.

[0098] The novelty, coincidence and attention parameters and particular software modules for the novelty-coincidence processing may be determined via manual trial and error. However, manual trial and error is a tedious and labor-intensive task. One preferred embodiment of the present invention applies the Genetic Algorithm (hereinafter referred to as "GA") to automatically determine an optimal set of parameters and modules. The GA is a very general method for optimization, well known to those in the art. It works by generating random variations and combinations from existing solutions, evaluating each variation in terms of some fitness function that is to be maximized, keeping the best solutions in a population, and applying recombination variation in a recursive procedure. The fitness function, for speech recognition, is some measure of accuracy of the entire algorithm, evaluated on a known and controlled set of speech.

[0099] In order to use the GA, the relevant parameters must be coded in a linear information array referred to herein as a "gene". In this case, we code the following general parameters:

1. Novelty parameters
2. Coincidence parameters and modules for each novelty output.
3. Attention parameters for each coincidence function.

One set of these general parameters that the GA generated is used in a preferred embodiment of the system 100, and are listed in FIGs. 4-9. The general parameters were optimized for each of the phonetic subgroups, vowels, fricatives, and nonfricative consonants. FIG. 4 shows the novelty parameters for fricatives. FIG. 4 shows six novelty outputs (channels). For each channel, FIG. 4 shows the center time-lag 204, the center length 206 and the center width 208, the scaling factor alpha, the surround time lag 212, the surround length 214 and the surround width 216. FIG. 5 shows the coincidence processing parameters and modules for fricatives. For each module, FIG. 5 shows the attention trigger 114, the time delta, the time start, the time stop, the frequency width (i.e., delta frequency) and the novelty channel upon which the module operates, all of which were generated by the GA. FIG. 6 shows the novelty parameters for vowels. FIG. 6 shows six novelty outputs (channels). For each channel, FIG. 6 shows the center time-lag 204,

the center length 206 and the center width 208, the scaling factor alpha, the surround time lag 212, the surround length 214 and the surround width 216. FIG. 7 shows the coincidence processing parameters and modules for vowels. For each module, FIG. 7 shows the attention trigger 114, the time delta, the time start, the time stop, the frequency width (i.e., delta frequency) and the novelty channel upon which the module operates, all of which were generated by the GA. FIG. 8 shows the novelty parameters for non-fricatives. FIG. 8 shows six novelty outputs (channels). For each channel, FIG. 8 shows the center time-lag 204, the center length 206 and the center width 208, the scaling factor alpha, the surround time lag 212, the surround length 214 and the surround width 216. FIG. 9 shows the coincidence processing parameters and modules for non-fricatives. For each module, FIG. 9 shows the attention trigger 114, the time delta, the time start, the time stop, the frequency width (i.e., delta frequency) and the novelty channel upon which the module operates, all of which were generated by the GA.

[0100] Each of these three processors (the novelty processor 102, the attention processor 112 and the coincidence processor 116) produces a vector of data values. The final output is a concatenation of all three vectors generated by these three processes.

[0101] FIG. 10 shows the specific data flow used in a preferred embodiment of the invention shown in FIG.1. FIG. 10 illustrates three similar processing units that run in parallel. A vowel processing unit 240 is optimized for detecting vowel phonics, a fricative processing unit 242 is optimized for fricatives, and a non-fricative processing unit 244 is optimized for non-fricative consonants. The outputs of the three processing units are collected (concatenated) into one large vector for the subsequent processing in the vector pattern recognizer 120. The used in the blocks of FIG. 10 correspond to the actual software module names in a preferred embodiment. FIG. 10 also denotes the details of data normalization and time-window formation.

[0102] The complete processing flow of the one preferred embodiment of the system 100 is specified by the high-level object-flow specification shown in FIGs. 11A and 11B. The terms used regarding the major steps in the object-flow specification are defined as follows:

Normalization - A pointwise mean and sigma normalization, based on fixed precomputed constant vectors, is done after the NoveltyRT and CoincidenceRT processes.

Extracting a Time-window - At every third input time (12msec), a 24-time window is selected from the normalized novelty data stream. This provides an appropriate time-frequency domain for coincidenceRT

Pattern Recognition - The outputs of the vowel, fricative, and nonfricative coincidence processes are all concatenated to form one large vector for each time. This vector is then applied to a vector classifier, or an array of phonetic detectors, as described in our previous patents.

Bayes Probabilities - In a manner similar to our previous patents, a non-parametric evaluation of the prior Bayes probabilities is performed offline. The log likelihood ratio curve is computed offline for each phoneme. At run-time, this curve is applied, at each time, and the log-likelihood estimates are sent on the search algorithm.

[0103] FIG. 12 shows a list of explanations for the modules, functions and parameters referred to in processing flow of FIGs. 11A and 11B.

[0104] FIGs. 13A and 13B, together show the ScaleMean software module used in the NoveltyRT software module from one preferred embodiment of the present invention.

[0105] FIG. 14A, 14B, 14C, 14D, 14E, 14F and 14G together show the NoveltyRT software module from one preferred embodiment of the present invention.

[0106] FIG. 15A, 15B, 15C, 15D, 15E, 15F, 15G, 15H and 15I together show the coincidenceRT and eTrigger software modules from one preferred embodiment of the present invention.

[0107] The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The present embodiments are therefore to be considered in respects as illustrative and not restrictive, the scope of the invention being indicated by the appended claims rather than by the foregoing description, and all changes which come within the meaning and range of the equivalency of the claims are therefore intended to be embraced therein.